A133153

ON THE FITTING OF PEARSON CURVES TO SUMS OF INDEPENDENT RANDOM VARIABLES

BY

THOMAS SELLKE

TECHNICAL REPORT NO. 333
MAY 19, 1983

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

DEPARTMENT OF STATISTICS
Stanford University
Stanford, California

ON THE FITTING OF PEARSON CURVES TO SUMS OF INDEPENDENT RANDOM VARIABLES

By

Thomas Sellke

Introduction and Summary.

In this report, we attempt to answer the following questions.

1. Is the sum of independent beta (Pearson Type I) random variables distributed as a beta random variable?

2. How well is the distribution of a sum of independent betas approximated by a beta distribution?

3. If two or more independent random variables are best fitted by one type of Pearson curve, is their sum best fitted by a Pearson curve of the same type?

Section 1 of this paper shows that the answer to the first question is "no". However, the calculations and computer simulations described in section 2 show that the sum of independent beta random variables often has a distribution which is close to a beta distribution, so that the answer to the second question is often "very well".

Section 3 shows that the answer to the third question depends on the Pearson curve type of the random variables and on whether they are identically distributed. Theorem 1 of this section shows that the sum of independent, identically distributed random variables of Pearson Type I, II, III or VII is best fitted by a Pearson curve of the same type. This is "almost" true for the other Pearson types in a certain sense. When the independent random variables to be added are not identically distributed, Pearson curve type is not preserved to this extent.

However, Theorem 2 and Theorem 3 can be used to determine the possible Pearson curve types of the sum given the Pearson curve type of the summands. There is some interest in the question of whether the sum of independent chi random variables is best fitted by a Pearson Type I distribution. (A single chi random variable is best fitted by a Pearson Type I.) The report finishes by showing that Pearson curves of Types I, III, IV, V, and VI can be best fitting for a sum of two independent chi random variables.

1. Is the Sum of Independent Beta (Pearson Type I) Random Variables Distributed as a Beta Random Variable?

It is easy to exhibit counterexamples, such as a sum of two independent $U[0,1]$ random variables. More generally, consider m independent betas with intervals of support $[0,a_1]$, $[0,a_2],\ldots,[0,a_m]$. It seems to be the case that the density of the sum of these betas will not be infinitely differentiable at points which can be written as the sum of some subset of the a_k's. Since the density of a beta is infinitely differentiable in the interior of its interval of support, this would imply that a sum of independent betas never has a beta distribution. A rigorous proof of this claim has not been worked out, however.

2. How Well is the Distribution of a Sum of Independent Betas Approximated by a Beta Distribution?

Percentage points were found for the Pearson curves whose first four moments agreed with the first four moments of various test distributions. These values are compared with the true percentage points or with percentage points obtained from computer simulation. The results are found in Tables 1-4. All the Pearson curves used were beta distributions.

Let U_1, U_2, and U_3 be independent U[0,1] random variables. Let $\beta_{2,2}$ and $\beta'_{2,2}$ be Beta(2,2) random variables independent of each other and of the U_i's. Table 1 gives percentage points, the Pearson curve approximations to these percentage points, and the true percentile values corresponding to the Pearson curve values for four symmetric test distributions. Note that the Pearson curve approximations do worst for $U_1 + U_2$, whose tent-shaped density does not look much like any beta density. The Pearson curves do about equally well for the other three test distributions.

Table 2 gives true, computer simulation, and Pearson curve percentage points for a sum of two $\beta(1,3)$ random variables. The computer simulation values were obtained by generating two independent random numbers uniformly distributed on [0,1], doing a transformation to obtain independent random numbers with a $\beta(1,3)$ distribution, recording the sum, and iterating this procedure 10^6 times. The other computer simulations were done in the same way, except that 5 and 10 independent $\beta(1,3)$ random numbers were added in each of the 10^6 iterations. The table shows that the computer simulation percentage points are in very good agreement with the true percentage points. The Pearson curve values are not as good, especially in the lower tail.

Tables 3 and 4 give computer simulation and Pearson curve percentage points for sums of 5 and 10 i.i.d. $\beta(1,3)$ random variables, respectively. The true percentage points were not found because the calculations would have been too messy, but Table 2 shows that the computer simulation values should be quite close to the true ones.

Table 4 also includes percentage points obtained from the Edgeworth expansion with Edgeworth correction terms of orders $n^{-1/2}$, n^{-1}, and $n^{-3/2}$. The different methods show very good agreement in both tables.

These results give an indication of how well the distribution of a sum of i.i.d. $\beta(p,q)$ random variables is approximated by a beta distribution when p and q are small positive integers. The Pearson curve approximation for a sum of two such betas gives only rough agreement with the true percentage points. One explanation of this behavior is that the density for a sum of two such betas exhibits a lack of "smoothness" at 1. For example, the "tent-function" density of $U_1 + U_2$ does not have a first derivative at 1, while the sum of two $\beta(1,3)$ random variables does not have a third derivative at 1. Thus, it is not surprising that such a density is not well approximated by a beta density, which is necessarily infinitely differentiable in its interval of support. As the number of iid betas which are added together increases, the smoothing effect of convolution on the density and the approach of the distribution toward the normal distribution makes the approximation by a beta better. Changing from integer values for p and q to real numbers of similar size should not seriously affect the quality of the approximations.

Moderate deviations from the identically distributed case should not make much difference either, although the next section will show that the Pearson curve which best fits a sum of independent, non-identically distributed betas is not always itself a beta. If p and

q are both very small positive numbers, it could be necessary to add a large number of these betas together before the sum distribution is smooth enough to be close to a beta. To take an extreme example, consider $p = q = 10^{-6}$. Such a $\beta(p,q)$ puts almost all of its mass very close to 0 or to 1. The distribution of a sum of k such betas would concentrate its mass close to the integers $0,1,2,\ldots,k$ unless k were quite large.

Table 1

True percentage points, Pearson curve approximations to these percentage points, and true percentiles for the Pearson curve vaues for four sum distributions.

	$U_1 + U_2$	$U_1 + U_2 + U_3$	$U_1 + \beta_{2,2}$	$\beta_{2,2} + \beta'_{2,2}$
Kurtosis	2.4	2.6	2.4107	2.5714
Range	[0,2]	[0,3]	[0,2]	[0,2]
True 0.25% point	.0707	.2466	.1390	.2112
Pearson value	.0348	.2318	.1331	.1990
True % for Pearson	.06%	.21%	.22%	.20%
True 0.5% point	.1000	.3107	.1763	.2536
Pearson value	.0789	.3077	.1737	.2512
True % for Pearson	.31%	.49%	.48%	.48%
True 1.0% point	.1414	.3915	.2242	.3052
Pearson value	.1342	.3966	.2241	.3056
True % for Pearson	.90%	1.04%	1.00%	1.00%
True 2.5% point	.2236	.5314	.3092	.3918
Pearson value	.2305	.5402	.3110	.3944
True % for Pearson	2.66%	2.63%	2.54%	2.56%
True 5.0% point	.3162	.6694	.3966	.4760
Pearson value	.3277	.6752	.3986	.4785
True % for Pearson	5.37%	5.13%	5.07%	5.09%
True 10.0% point	.4472	.8434	.5123	.5824
Pearson value	.4554	.8428	.5133	.5832
True % for Pearson	10.37%	9.98%	10.05%	10.04%
25.0%	.7071	1.1471	.7338	.7761
Pearson value	.7003	1.1452	.732	.775
True % for Pearson	24.52%	24.88%	24.90%	24.85%

Table 2

True, computer simulation (10^6 rep.), and Pearson curve percentage points for a sum of two $\beta(1,3)$ r.v.'s.

Percent	True	Computer Simulation	Pearson Curve	True Percentiles of Pearson Curve Values
.25%	0.0240	0.0243	0.0071	.022%
.5%	0.0341	0.0343	0.0210	.19%
1.0%	0.0487	0.0492	0.0396	.67%
2.5%	0.0786	0.0790	0.0748	2.28%
5%	0.1139	0.1142	0.1137	4.98%
10%	0.1672	0.1676	0.1696	10.26%
25%	0.2884	0.2886	0.2909	25.34%
50%	0.4669	0.4667	0.4656	49.83%
75%	0.6766	0.6765	0.6743	74.78%
90%	0.8769	0.8770	0.8796	90.14%
95%	1.000	1.0001	1.0048	95.14%
97.5%	1.1091	1.1089	1.1122	97.55%
99%	1.2353	1.2358	1.2333	98.98%
99.5%	1.3187	1.3183	1.3123	99.47%
99.75%	1.3930	1.3910	1.3824	99.72%

Table 3

Percentage points for a sum of 5 iid $\beta(1,3)$ random variables

Percent	Computer Simulation	Pearson Curve
.25%	0.287	0.279
.5%	0.338	0.326
1.0%	0.396	0.395
2.5%	0.497	0.496
5%	0.588	0.592
10%	0.709	0.712
25%	0.938	0.938
50%	1.221	1.220
75%	1.531	1.531
90%	1.828	1.828
95%	2.010	2.012
97.5%	2.170	2.173
99%	2.358	2.362
99.5%	2.488	2.490
99.75%	2.610	2.609

Table 4

Percentage points for a sum of 10 iid $\beta(1,3)$ random variables.

Percent	Computer Simulation	Pearson Curve	Edgeworth Expansion
.25%	1.0112	1.0092	1.0157
.5%	1.1068	1.1070	1.1084
1.0%	1.2160	1.2172	1.2160
2.5%	1.3860	1.3881	1.3859
5%	1.5416	1.5434	1.5417
10%	1.7308	1.7319	1.7311
25%	2.0680	2.0683	2.0685
50%	2.4724	2.4710	2.4715
75%	2.9004	2.9006	2.9005
90%	3.3040	3.3060	3.3055
95%	3.5520	3.5555	3.5553
97.5%	3.7688	3.7752	3.7752
99%	4.0256	4.0337	4.0332
99.5%	4.2056	4.2112	4.2100
99.75%	4.3716	4.3764	4.3750

3. If Two or More Independent Random Variables are Best Fitted by One Type of Pearson Curve, is Their Sum Best Fitted by a Pearson Curve of the Same Type?

The answer to this question will depend on the type of Pearson curve which best fits the summand random variables. However, before the investigation of this question can begin, it will be necessary to establish some notation and to make some background remarks concerning the Pearson curve system.

Let X_1 and X_2 be independent random variables with finite fourth moments. Let K_1, K_2, K_3, and K_4 be the first four cumulants of X_1. Let L_1, L_2, L_3, and L_4 be the first four cumulants of X_2. The first four cumulants of $X_1 + X_2$ will be $K_1 + L_1$, $K_2 + L_2$, $K_3 + L_3$, and $K_4 + L_4$. Let $\sqrt{\beta_1'}$, $\sqrt{\beta_1''}$, and $\sqrt{\hat{\beta}_1}$ be the skewness values for X_1, X_2, and X_1+X_2 respectively. Let β_2', β_2'', and $\hat{\beta}_2$ be the kurtosis values for X_1, X_2, and X_1+X_2, respectively. Recall that $\sqrt{\beta_1'}$ and β_2' are defined by

$$\sqrt{\beta_1'} = \frac{K_3}{K_2^{3/2}} \qquad \text{and} \qquad \beta_2' = 3 + \frac{K_4}{K_2^2} .$$

The other skewness and kurtosis values are defined analogously. The symbols $\sqrt{\beta_1}$ and β_2 will be used as generic symbols for skewness and kurtosis.

A Pearson curve is uniquely determined by its first four moments. Thus, a natural way to fit a Pearson curve to a probability distribution is to find the Pearson curve whose first four moments match those of

the distribution. In this discussion, the "best fitting" Pearson curve will be defined to be the one found in this way. However, other fitting methods are sometimes used. For example, Pearson curves are sometimes fitted to chi random variables so as to match the first three moments subject to the constraint that 0 be the left endpoint of the interval of support.

Up to location and scale, the Pearson curve which best fits a distribution is determined by the skewness $\sqrt{\beta_1}$ and the kurtosis β_2 of the distribution. Since the type of a Pearson curve is location and scale invariant, $\sqrt{\beta_1}$ and β_2 determine the type.

The following formulas, taken from Johnson and Kotz (1970), show how to find Pearson curve type from $\sqrt{\beta_1}$ and β_2. Define c_0, c_1, c_2, and κ by

$$c_0 = (4\beta_2 - 3\beta_1)(10\beta_2 - 12\beta_1 - 18)^{-1}\mu_2$$

$$c_1 = \sqrt{\beta_1}\,(\beta_2 + 3)(10\beta_2 - 12\beta_1 - 18)^{-1}\sqrt{\mu_2}$$

$$c_2 = (2\beta_2 - 3\beta_1 - 6)(10\beta_2 - 12\beta_1 - 18)^{-1}$$

$$\kappa = \frac{1}{4}\, c_1^2\, (c_0 c_2)^{-1} .$$

Type I: $\kappa < 0$, which is equivalent to $2\beta_2 - 3\beta_1 - 6 < 0$.

Type II: $\beta_1 = 0$, $\beta_2 < 3$.

Type III: $2\beta_2 - 3\beta_1 - 6 = 0$.

Type IV: $0 < \kappa < 1$.

Type V: $\kappa = 1$.

Type VI: $\kappa > 1$.

Type VII: $\beta_1 = 0, \quad \beta_2 > 3$.

The classification of (β_1,β_2) pairs implied by these formulas is displayed graphically on the next two pages, which are taken from Rhind (1909). The "limit for all frequency distributions" line has been added to Rhind's version of Figure 1. The text of Rhind's paper indicates that existence of this limiting line was not known in 1909. The line labeled V in Figure 1 may look like it is not quite straight because of sloppiness on Rhind's part, but this is not the case. This curve is the solution to the cubic equation

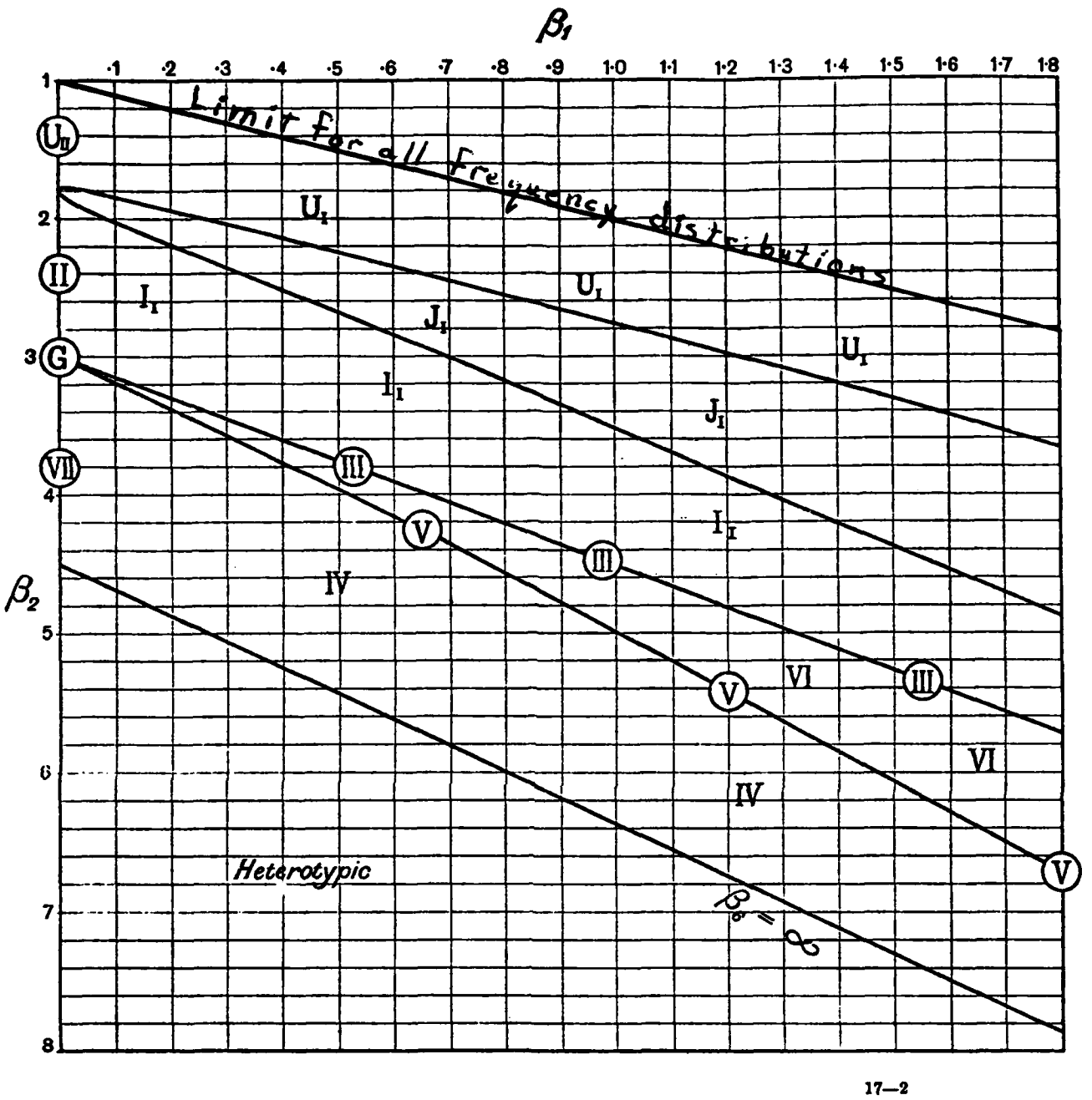
$$\beta_1(\beta_2+3)^2 = 4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6) .$$

The curve is also shown on Figure 2, where it is more obvious that it is not straight. The line labeled III is straight, however,

The kurtosis β_2 does not seem to be a convenient parameter for the purposes of this discussion. For this reason, let us define γ', γ'', and $\hat{\gamma}$ by

$$\gamma' = \beta_2'-3 , \quad \gamma'' = \beta_2''-3 , \quad \text{and} \quad \hat{\gamma} = \hat{\beta}_2-3 .$$

Thus, the γ parameters are related to the cumulants by



This diagram, taken from Rhind (1909), shows how β_1 and β_2 determine Pearson curve type.

"U-shaped" betas fall in U_I.

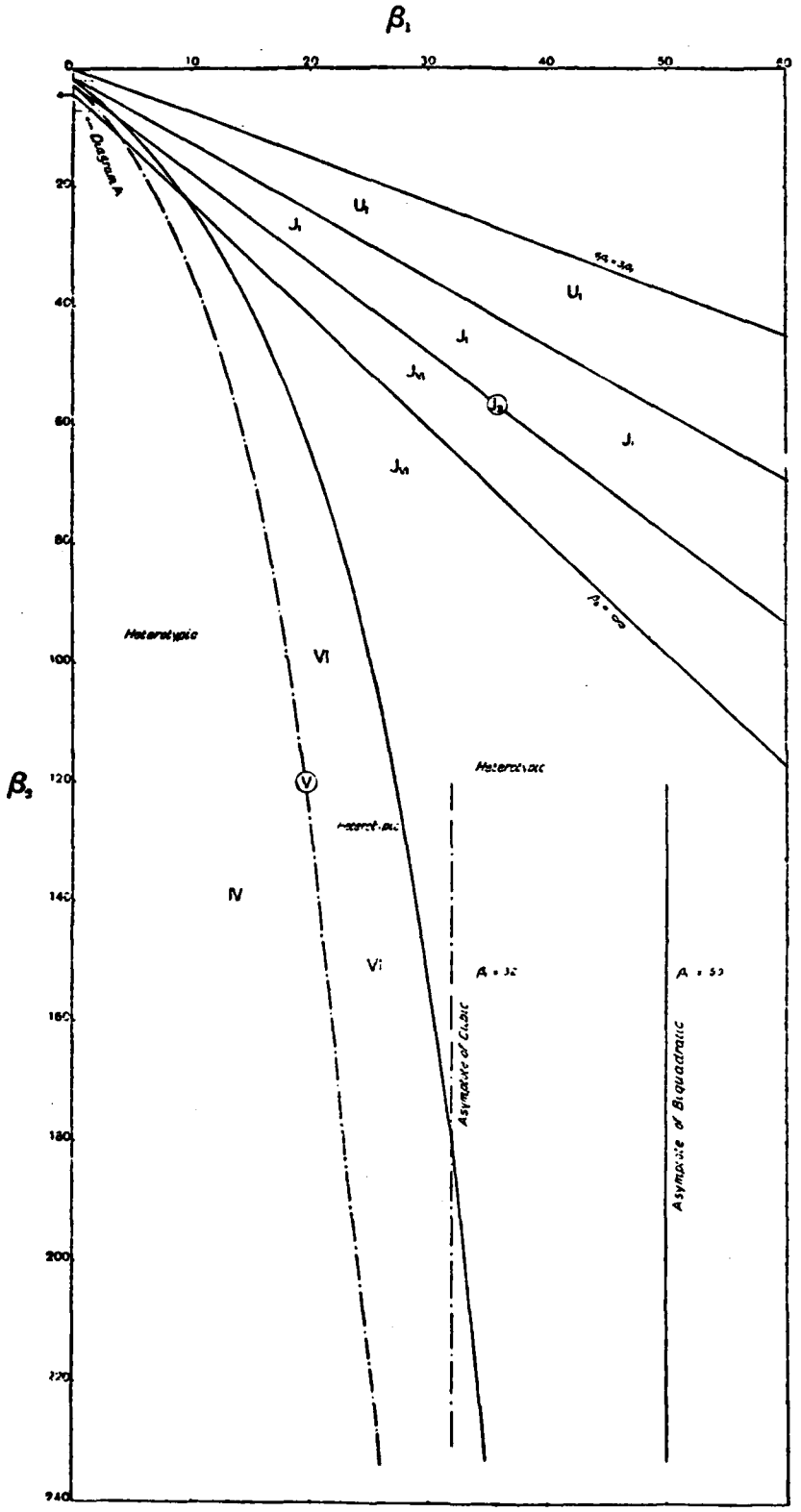
"J-shaped" betas fall in J_I.

Other betas fall in I_I.

For all distributions, (β_1,β_2) satisfies $\beta_2-\beta_1-1 < 0$.

Pearson curves for which (β_1,β_2) falls below the "$\beta_6 = \infty$" line have an infinite 8th moment.

Figure 1.



This diagram, taken from Rhind (1909), relates Pearson curve type to β_1 and β_2 for a larger part of the (β_1,β_2) plane than is covered by Figure 1.

Figure 2.

$$\gamma' = \frac{K_4}{K_2^2}, \quad \gamma'' = \frac{L_4}{L_2^2}, \quad \hat{\gamma} = \frac{K_4+L_4}{(K_2+L_2)^2}.$$

One can think of this γ parameter as being a normalized fourth cumulant in the same way that $\sqrt{\beta_1}$ is a normalized third cumulant.

The Pearson curve corresponding to a given distribution is of course specified, up to location and scale, by the values of $\sqrt{\beta_1}$ and γ of the distribution. When one works in terms of β_1 and γ instead of in terms of β_1 and β_2, Figure 1 is replaced by Figure 3.

Let us subdivide the region in the (β_1,γ) plane which corresponds to Type I distributions into the regions I^-, I^+, and I^0. (See Figure 4.) I^- is the part of the Type I region where $\gamma < 0$, I^+ is the part of the Type I region where $\gamma > 0$, and I^0 is the part of the Type I region where $\gamma = 0$. These subregions have no known significance with respect to the shapes of the Pearson curves they contain. Their importance arises solely from the question to be investigated.

If $\beta_1 = 0$, the Pearson curve type is determined by the sign of γ:

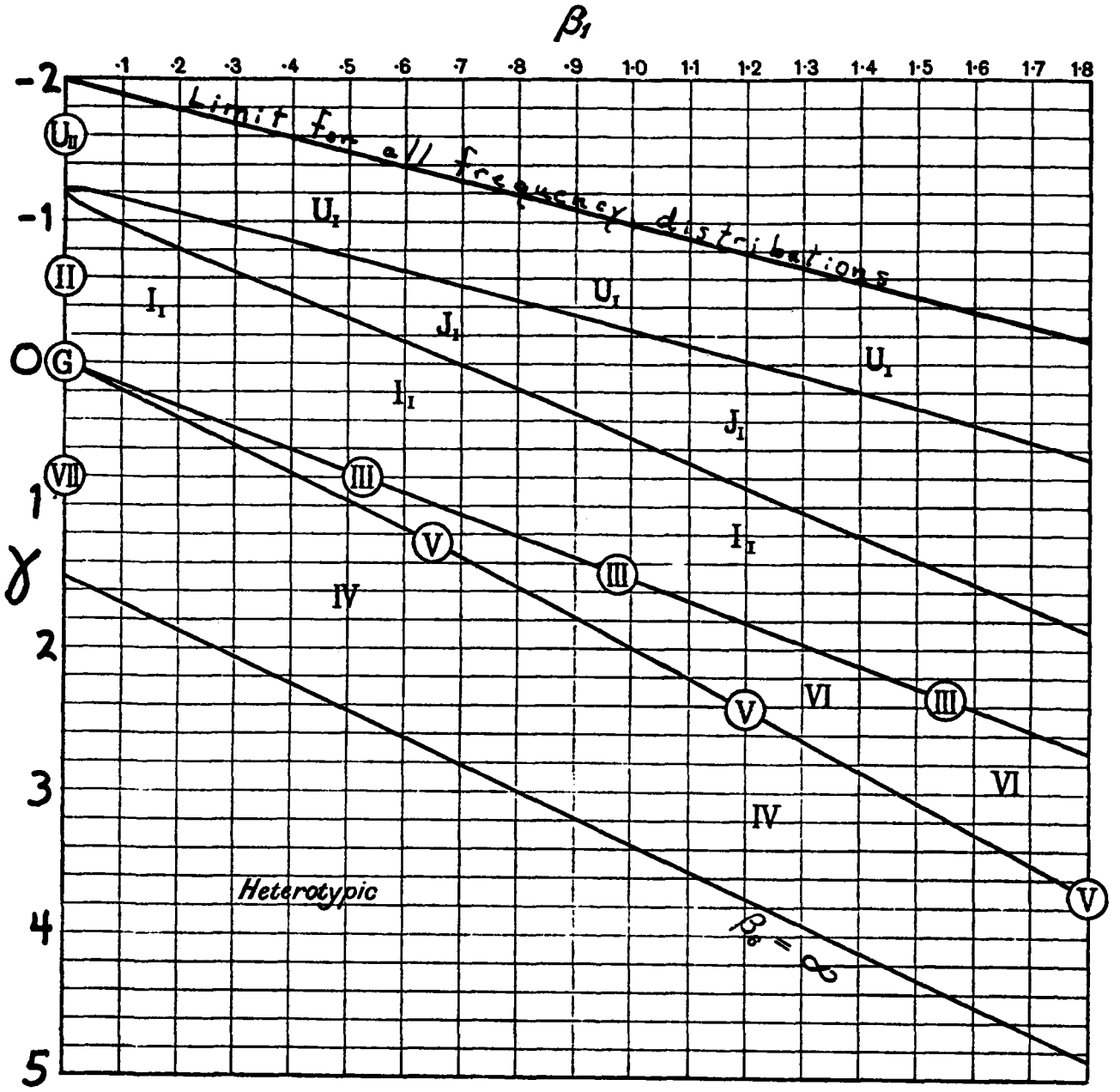
$\beta_1 = 0, \quad \gamma < 0$ implies Type II.

$\beta_1 = 0, \quad \gamma = 0$ implies Type G (normal distribution)

$\beta_1 = 0, \quad \gamma > 0$ implies Type VII.

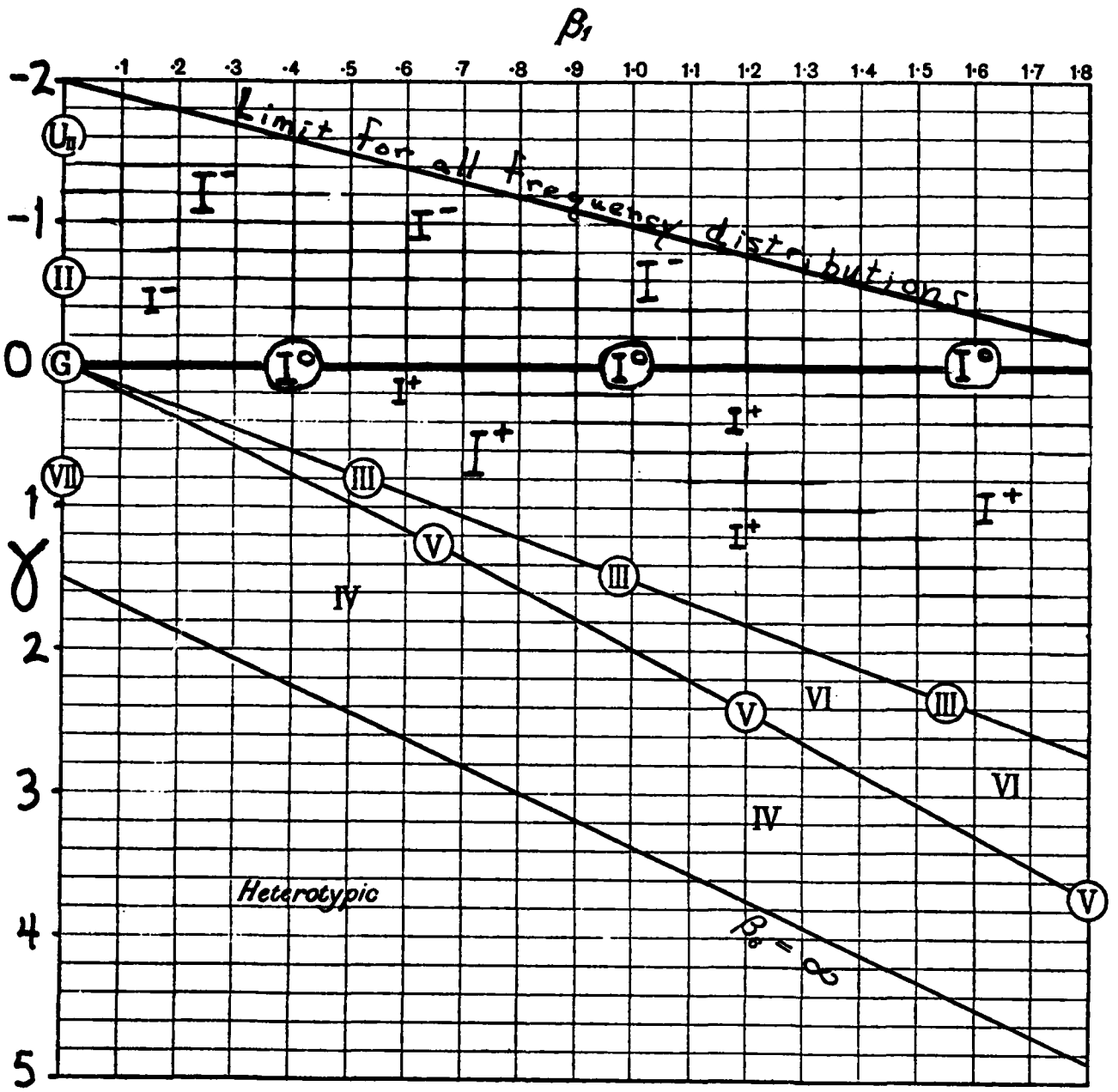
If $\beta_1 > 0$, the Pearson curve type is "almost" determined by the ratio $\frac{\gamma}{\beta_1}$:

$\frac{\gamma}{\beta_1} < 0$ implies Type I^-.



This diagram is the same as Figure 1, except that the vertical axis is parameterized by γ instead of β_2.

Figure 3.



This diagram, taken from Rhind (1909), shows how the Type I region in the (β_1,γ) plane is divided into the regions I^- and I^+ and the line I^0.

Figure 4.

$\frac{\gamma}{\beta_1} = 0$ implies Type I^0.

$0 < \frac{\gamma}{\beta_1} < \frac{3}{2}$ implies Type I^+.

$\frac{\gamma}{\beta_1} = \frac{3}{2}$ implies Type III.

$\frac{3}{2} < \frac{\gamma}{\beta_1}$ implies Type VI, Type V, or Type IV.

Now if one restricts attention to that part of the (β_1,γ) plane shown in Figure 4, there exists some small number $\varepsilon > 0$ such that

$\frac{3}{2} < \frac{\gamma}{\beta_1} < 2-\varepsilon$ implies Type VI (true even when $\beta_1 > 1.8$),

$2-\varepsilon \leq \frac{\gamma}{\beta_1} \leq 2+\varepsilon$ is implied by Type V,

$2+\varepsilon < \frac{\gamma}{\beta_1}$ implies Type IV.

This completes the necessary background remarks, so that we can finally procede to the question of interest. To begin, let us consider what happens when X_1 and X_2 are iid, or, to restrict attention to what is relevant here, when X_1 and X_2 are such that $K_2 = L_2$, $K_3 = L_3$, and $K_4 = L_4$. In this case, we have

$$\hat{\gamma} = \frac{K_4+L_4}{(K_2+L_2)^2} = \frac{2K_4}{4K_2^2} = \frac{\gamma'}{2} = \frac{\gamma''}{2} .$$

$$\hat{\beta}_1 = \frac{(K_3+L_3)^2}{(K_2+L_2)^3} = \frac{4K_3^2}{8K_2^3} = \frac{\beta_1'}{2} = \frac{\beta_1''}{2} .$$

If $\beta_1' = 0$, this implies $\hat{\beta}_1 = 0$ and $\text{sign}(\hat{\gamma}) = \text{sign}(\gamma')$. If $\beta_1' > 0$, this implies $\frac{\hat{\gamma}}{\beta_1} = \frac{\gamma'}{\beta_1'}$.

Thus, the Types II, G, and VII, which occur when $\beta_1 = 0$, are preserved under addition of two iid random variables. The same is true of Types I^-, I^o, I^+, and III, which are characterized by the value of the ratio $\frac{\gamma}{\beta_1}$, since this ratio is preserved under addition of two iid random variables. The Types VI, V, and IV are "almost" preserved in the same sense that they are "almost" determined by the ratio $\frac{\gamma}{\beta_1}$. Thus, Type VI random variables for which $3/2 < \frac{\gamma}{\beta_1} < 2-\epsilon$ are preserved under addition in this sense. The same is true for Type IV random variables for which $2+\epsilon < \frac{\gamma}{\beta_1}$ and $0 < \beta_1 \leq 1.8$. Type V random variables will almost never be preserved, but the sum of two iid Type V's for which $0 < \beta_1 \leq 1.8$ will be very close to a Type V distribution with respect to its first four moments. The second derivative of the Type V curve is negative close to $\beta_1 = 0$ and is positive when β_1 is large, so there will be at least one point (β_1,γ) on the curve for which $(\frac{\beta_1}{2}, \frac{\gamma}{2})$ is also on the Type V curve.

If n iid random variables are added, the β_1 and γ values for the sum are equal to $\frac{1}{n}$ times the corresponding values for the summands. It follows that all of the above results hold when n instead of just two iid random variables are added together. Let us record this formally as

Theorem 1. Suppose a random variable X is best fitted by a Pearson curve of Type I, II, III, or VII. If n iid copies of X are added together, the sum is best fitted by a Pearson curve of the same type. The same is "almost" true for Pearson curve Types IV, V, and VI in the sense described above.

When X_1 and X_2 are not iid, matters become more complicated. The relationship between the (β_1', γ') and (β_1'', β'') pairs of the summands and the $(\hat{\beta}_1, \hat{\gamma})$ pair of the sum is not so easily described as in the iid case. The key result here will be Theorem 3, although Theorem 2 will be useful also.

Theorem 2. $\hat{\beta}_1 \leq \max\{\beta_1', \beta_1''\}$, and $|\hat{\gamma}| \leq \max\{|\gamma'|, |\gamma''|\}$.

Proof. Suppose $\beta_1'' \leq \beta_1'$. In terms of the cumulants, this means

$$\frac{L_3^2}{L_2^3} \leq \frac{K_3^2}{K_2^3},$$

so that

$$|L_3| \leq |K_3| \left(\frac{L_2}{K_2}\right)^{3/2}.$$

Thus,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{(K_3+L_3)^2}{(K_2+L_2)^3} \\
&\leq \frac{(|K_3|+|L_3|)^2}{(K_2+L_2)^3} \\
&\leq \frac{\left(|K_3|+|K_3|\left(\frac{L_2}{K_2}\right)^{3/2}\right)^2}{(K_2+L_2)^3}
\end{aligned}$$

$$\leq \frac{K_3^2}{K_2^3} \cdot \frac{(K_2^{3/2} + L_2^{3/2})^2}{(K_2+L_2)^3}$$

$$\leq \beta_1' \left[\frac{K_2^{3/2} + L_2^{3/2}}{(K_2+L_2)^{3/2}} \right]^2 .$$

Since $x^{3/2}$ is a convex increasing function of x for $x \geq 0$,

$$K_2^{3/2} + L_2^{3/2} \leq (K_2+L_2)^{3/2} .$$

This and the above imply

$$\hat{\beta}_1 \leq \beta_1' = \max\{\beta_1', \beta_1''\} .$$

The proof of the second assertion is similar. ∎

Theorem 3. If γ' and γ'' have the same sign (positive, negative, or 0), then $\hat{\gamma}$ also has this sign, and

$$\left|\frac{\hat{\gamma}}{\hat{\beta}_1}\right| \geq \min\left\{\left|\frac{\gamma'}{\beta_1'}\right| , \left|\frac{\gamma''}{\beta_1''}\right|\right\} .$$

Here, $\left|\frac{a}{0}\right|$ is to be interpreted as ∞ for every $a \in \mathbb{R}$. If the sign of γ' and γ'' is not 0, then equality holds if and only if

$$\frac{K_3}{K_2} = \frac{L_3}{L_2} \quad \text{and} \quad \frac{K_4}{K_2} = \frac{L_4}{L_2} .$$

Proof. The proof is trivial when $\text{sign}(\gamma') = \text{sign}(\gamma'') = 0$.

Suppose that $\gamma' > 0$ and $\gamma'' > 0$. This implies $\hat{\gamma} > 0$.

Note that $\beta_1 = (\sqrt{\beta_1})^2$ is never negative. Thus, the assertion is equivalent to

$$\frac{\hat{\beta}}{\hat{\gamma}} \leq \max\{\frac{\beta_1'}{\gamma'} \quad \frac{\beta_1''}{\gamma''}\} \ .$$

Let

$$c = \max\{\frac{\beta_1'}{\gamma'} \ , \quad \frac{\beta_1''}{\gamma''}\} \ .$$

Then

(1) $$\frac{\hat{\beta}_1}{\hat{\gamma}} \leq c$$

if and only if

$$\hat{\beta}_1 \leq c\hat{\gamma}$$

if and only if

$$\frac{(K_3+L_3)^2}{(K_2+L_2)^3} \leq c \frac{K_4+L_4}{(K_2+L_2)^2}$$

if and only if

$$(K_3+L_3)^2 \leq c(K_4+L_4)(K_2+L_2)$$

if and only if

$$(2) \qquad (K_2^{3/2}\sqrt{\beta_1'} + L_2^{3/2}\sqrt{\beta_1''})^2 \leq c(K_2^2\gamma' + L_2^2\gamma'')(K_2+L_2)$$

if

$$(3) \qquad (K_2^{3/2}\sqrt{\beta_1'} + L_2^{3/2}\sqrt{\beta_1''})^2 \leq (K_2^2\beta_1' + L_2^2\beta_1'')(K_2+L_2)$$

if and only if

$$K_2^3\beta_1' + L_2^3\beta_1'' + 2K_2^{3/2}L_2^{3/2}\sqrt{\beta_1'}\sqrt{\beta_1''} \leq K_2^3\beta_1' + L_2^3\beta_1'' + L_2K_2^2\beta_1' + K_2L_2^2\beta_1''$$

if and only if

$$2K_2^{3/2}L_2^{3/2}\sqrt{\beta_1'}\sqrt{\beta_1''} \leq L_2K_2^2\beta_1' + K_2L_2^2\beta_1''$$

if and only if

$$2K_2^{1/2}L_2^{1/2}\sqrt{\beta_1'}\sqrt{\beta_1''} \leq K_2\beta_1' + L_2\beta_1''$$

if and only if

$$(4) \qquad 0 \leq (K_2^{1/2}\sqrt{\beta_1'} - L_2^{1/2}\sqrt{\beta_1''})^2 \quad .$$

Equation (4) is always true. Following the chain of implications back up shows equation (1) is always true.

If $K_2^{1/2}\sqrt{\beta_1'} \neq L_2^{1/2}\sqrt{\beta_1''}$, then we have strict inequality in (4). Strict inequality in (4) implies strict inequality in all the preceding steps. Note that $K_2^{1/2}\sqrt{\beta_1'} = L_2^{1/2}\sqrt{\beta_1''}$ is equivalent to $\frac{K_3}{K_2} = \frac{L_3}{L_2}$. If $\frac{\beta_1'}{\gamma'} \neq \frac{\beta_1''}{\gamma''}$, then we get strict inequality in (2) when we go up from (3) to (2). Strict inequality in (2) implies strict inequality in (1). Note that $\frac{K_4}{K_2} = \frac{L_4}{L_2}$ is equivalent to $\frac{\beta_1'}{\gamma'} = \frac{\beta_1''}{\gamma''}$ when $\frac{K_3}{K_2} = \frac{L_3}{L_2}$ holds. This shows that the "only if" part of the last assertion.

If $\frac{K_3}{K_2} = \frac{L_3}{L_2}$ and $\frac{K_4}{K_2} = \frac{L_4}{L_2}$ both hold, then $K_2^{1/2}\sqrt{\beta_1'} = L_2^{1/2}\sqrt{\beta_1''}$ and $\frac{\beta_1'}{\gamma'} = \frac{\beta_1''}{\gamma''}$. This implies inequality in (4) and in all the preceding steps. This finishes the case $\gamma' > 0$ and $\gamma'' > 0$. The proof for the case $\gamma' < 0$ and $\gamma'' < 0$ is similar. ▨

It follows from Theorem 3 that if independent random variables are best fitted by Type II Pearson curves, then their sum is also best fitted by a Type II Pearson curve. The same holds for Type VII, for the union of Type I^o and Type G, and for the union of Type I^- and Type II. These results would have been trivial to prove directly, however. It is on the types for which $\gamma > 0$ that Theorem 3 sheds the most light. For example, if (β_1', γ') and (β_1'', γ'') are both in region I^+, then $(\hat{\beta}_1, \hat{\gamma})$ may fall only in I^+, III, VI, V, IV, and VII. However, if (β_1', γ') and (β_1'', γ'') fall in the Type VI region, then $(\hat{\beta}_1, \hat{\gamma})$ must be in VI, V, IV, or VII. By using both Theorem 2

and Theorem 3, one can conclude that $(\hat{\beta}_1,\hat{\gamma})$ will be in either IV or VII when (β_1',γ') and (β_1'',γ'') are in that part of the IV region for which $0 < \beta_1 \leq 1.8$ and $2+\varepsilon < \frac{\gamma}{\beta_1}$.

The most interesting application of Theorem 3 is to Type III random variables. Suppose that X_1 and X_2 both have gamma distributions with support on $[0,\infty)$. Then the densities of X_1 and X_2 are Type III Pearson curves, so that (β_1',γ') and (β_1'',γ'') fall on the line $\frac{\gamma}{\beta_1} = \frac{3}{2}$. By the first part of Theorem 3, $(\hat{\beta}_1,\hat{\gamma})$ must fall in III, VI, V, IV, or VII. However, the fact that X_1 and X_2 are gamma distributions with right tails implies $\sqrt{\beta_1'} > 0$ and $\sqrt{\beta_1''} > 0$. This in turn implies $\sqrt{\hat{\beta}_1} > 0$, so that $(\hat{\beta}_1,\hat{\gamma})$ will not fall in VII. Now we can apply the condition for equality in Theorem 3. When X_1 and X_2 are both gamma random variables, the condition for equality in Theorem 3 is equivalent to the condition that the scale parameters of X_1 and X_2 be the same. Two gamma random variables with the same scale parameter are, at least in a limiting sense, sums of iid copies of the same random variable. (Recall that a gamma random variable with shape parameter k and scale parameter λ can be thought of as a sum of k independent exponential random variables with parameter λ. This interpretation is useful even when k is not a integer.) Thus, we are essentially back in the case covered by Theorem 1 when the equality condition of Theorem 3 holds for gamma random variables. On the other hand, Theorem 3 implies that X_1+X_2 will have the first four moments of a Pearson curve of Types VI, V, or IV when X_1 and X_2 have different scale parameters. Thus, the sum of two gamma random variables with different scale parameters cannot have the first four moments of a gamma random variable.

It may also be enlightening to look more closely at what can happen when (β_1',γ') and (β_1'',γ'') are in I^+. By Figure 4, it is possible for X_1 to have a beta distribution such that $(\beta_1',\gamma') = (1,1)$. Suppose this holds if $X_1 \sim \beta(p,q)$, and that $\sqrt{\beta_1'} = 1$. Suppose further that $X_2 \sim \beta(q,p)$. Then $(\beta_1'',\gamma'') = (1,1)$, but $\sqrt{\beta_1''} = -1$. Note also that, modulo a location shift, X_2 will have the same distribution as $-X_1$. In this case, we will have $(\hat{\gamma},\hat{\beta}_1) = (\frac{1}{2},0) \in$ VII. Thus, the sum of two beta random variables can have the same first four moments as a t distribution. This will be the case whenever $(\beta_1',\gamma') \in I^+$ and X_2 has the same distribution as $-X_1$.

Now suppose that $(\beta_1',\gamma') = (1,1)$, and that $(\beta_1'',\gamma'') = (0,0)$. Thus, X_2 will have the same first four moments as a normal distribution. Calculation of $\hat{\beta}_1$ and $\hat{\gamma}$ yields

$$\hat{\beta}_1 = \frac{(K_3+L_3)^2}{(K_2+L_2)^2} = \frac{K_3^2}{(K_2+L_2)^3}$$

and

$$\hat{\gamma} = \frac{K_4+L_4}{(K_2+L_2)^2} = \frac{K_4}{(K_2+L_2)^2},$$

since $L_3 = L_4 = 0$. Since $\beta_1' = \frac{K_3^2}{K_2^3}$ and $\gamma' = \frac{K_4}{K_2^2}$, we have

$$\frac{\hat{\gamma}}{\hat{\beta}_1} = \frac{\gamma'}{\beta_1'}\,\frac{K_2+L_2}{K_2} = \frac{K_2+L_2}{K_2}.$$

Now K_2 and L_2 can be chosen independently of (β_1',γ') and (β_1'',γ'').

This is true because K_2 is just the variance of X_1, so that K_2 can be varied by scale transformations which leave (β_1', γ') unchanged. The same holds for L_2 of course. Thus, by properly choosing K_2 and L_2, $\frac{\hat{\gamma}}{\hat{\beta}} = \frac{K_2+L_2}{K_2}$ can be made equal to any given number in $(1,\infty)$. This result and Theorem 2 imply that $(\hat{\gamma},\hat{\beta}_1)$ can be made to fall into any of I^+, III, VI, V, and IV in this case. Since $\hat{\beta}_1$ and $\hat{\gamma}$ are continuous functions of the cumulants of X_1 and X_2, it is not hard to see that $(\hat{\beta}_1,\hat{\gamma})$ can fall into any of I^+, III, VI, V, and IV even when (β_1', γ') and (β_1'', γ'') are in I^+ and $\sqrt{\beta_1'}$ has the same sign as $\sqrt{\beta_1''}$.

Interest has been expressed in the fitting of Pearson curves to sums of independent chi random variables. Results contained in Elandt (1961) are helpful here. The Elandt paper gives formulas for the moments of noncentral chi random variables. It also contains a diagram (Figure 1, p. 555) showing how the (β_1,β_2) pair for a noncentral chi moves through the (β_1,β_2) plane as the noncentrality parameter changes. Comparison of this diagram with Figure 1 on page 13 of this paper shows that a chi random variable is always best fitted by a Type I Pearson curve. By Theorem 1, any sum of finitely many iid chi random variables is also best fitted by a Type I Pearson curve. The question of whether this is true for nonidentically distributed summands now arises. The following shows that the best fitting Pearson curve for the sum of a central chi random variable and an independent noncentral chi random variable can be of Type I, III, IV, V, or VI.

Let X_1 be a central chi random variable arising from taking the absolute value of a $N(0,1)$. Let X_2 be a $N(0,1)$ random variable independent of X_1. It will now be shown that X_1+X_2 has the first four moments of a Type IV Pearson curve.

It is easy to find the first four cumulants K_1, K_2, K_3, and K_4 of X_1 from the first row of Table 1 in Elandt (1961). The calculations imply

$$K_1 = 0.7979$$

$$K_2 = 0.3634$$

$$K_3 = 0.21804$$

$$K_4 = 0.11473 .$$

The first four cumulants of X_2 are of course

$$L_1 = 0$$

$$L_2 = 1.0$$

$$L_3 = 0$$

$$L_4 = 0 .$$

If we again use $\hat{\beta}_1$ and $\hat{\beta}_2$ for the (skewness)^2 and kurtosis of X_1+X_2, we get

$$\hat{\beta}_1 = \frac{(K_3+L_3)^2}{(K_2+L_2)^3} = \frac{(0.21804)^2}{(1.3634)^3} = 0.01876$$

and

$$\hat{\beta}_2 = 3 + \frac{K_4+L_4}{(K_2+L_2)^2} = 3 + \frac{0.1174}{(1.3634)^2} = 3.06172 .$$

Also, $\hat{\gamma} = \hat{\beta}_2 - 3 = 0.06174$. Thus, $\frac{\hat{\gamma}}{\hat{\beta}_1} = 3.290$. It is easy to see from Figure 3 that $(\hat{\beta}_1, \hat{\gamma}) \in$ IV.

Since the second, third, and fourth cumulants of a chi arising from $N(\mu,1)$ approach those of a normal $N(0,1)$ as $\mu \to \infty$, the continuity of $\hat{\beta}_1$ and $\hat{\gamma}$ as functions of the cumulants implies that the sum of the central chi $|N(0,1)|$ and the noncentral chi $|N(\mu,1)|$

will have $(\hat{\beta}_1,\hat{\gamma})$ in IV for μ sufficiently large. As μ varies from 0 to ∞, the $(\hat{\beta}_1,\hat{\gamma})$ pair will trace out a continuous curve in the (β_1,γ) plane which starts in the Type I region and ends in the Type IV region. By continuity and the fact that $\hat{\beta}_1$ is positive everywhere along this curve, the curve must pass through the regions for Types I, III, VI, V, and IV. Calculations using moments for $|N(3,1)|$ obtained from the last row of Table 1 in Elandt (1961) show that $(\hat{\beta}_1,\hat{\gamma})$ is still in I^+ when $\mu = 3$.

References

Elandt, Regina C. (1961). The folded normal distribution: Two methods of estimating parameters from moments. Technometrics, 3, 551-562.

Johnson, N.L. and Kotz, S. (1970). Continuous Univariate Distributions - 1, New York: Houghton Mifflin.

Rhind, A. (1909). Tables to facilitate the computation of probable errors of the chief constants of skew frequency distributions. Biometrika, 7, 127-147.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 333	2. GOVT ACCESSION NO.: AD-A133 153	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): ON THE FITTING OF PEARSON CURVES TO SUMS OF INDEPENDENT RANDOM VARIABLES		5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): THOMAS SELLKE		8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS: DEPT. OF STATISTICS STANFORD UNIVERSITY - STANFORD, CALIF.		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS: STATISTICS & PROBABILITY PROGRAM (code 411(SP) OFFICE OF NAVAL RESEARCH ARLINGTON, VA. 22217		12. REPORT DATE: MAY 19, 1983 13. NUMBER OF PAGES: 30
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Beta distribution, Pearson curves, sums of independent random variables.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE REVERSE SIDE

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT NO. 333

ON THE FITTING OF PEARSON CURVES TO SUMS OF INDEPENDENT RANDOM VARIABLES

By

Thomas Sellke

ABSTRACT

It is shown that the distribution of a sum of independent beta random variables is often well approximated by a properly scaled beta distribution. The relationship between the type of Pearson curve which best fits a sum of independent random variables and the types of the Pearson curves which best fit the summand random variables is also investigated. The best fitting Pearson curve for a distribution is defined here to be the unique Pearson curve with the same first four moments.